AFRL-RH-WP-TR-2012-0165

# FOREIGN LANGUAGE ANALYSIS AND RECOGNITION (FLARE) INITIAL PROGRESS

**Brian M. Ore**
**Stephen A. Thorn**
**David M. Hoeferlin**
**SRA International**
**5000 Springfield Street, Suite 200**
**Dayton, OH 45431**

**Raymond E. Slyh**
**Eric G. Hansen**
**Air Force Research Laboratory**
**711th Human Performance Wing**
**Human Effectiveness Directorate**
**Human Trust and Interaction Branch**
**2255 H St.**
**Wright-Patterson AFB, OH 45433**

**November 2012**
**Interim Report for 1 October 2010 to 30 September 2012**

**Distribution A: Approved for public release; distribution is unlimited.**

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
HUMAN EFFECTIVENESS DIRECTORATE
HUMAN-CENTERED ISR DIVISION
HUMAN TRUST AND INTERACTION BRANCH
WRIGHT-PATTERSON AFB OH 45433
AIR FORCE MATERIAL COMMAND
UNITED STATES AIR FORCE

# NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

**AFRL-RH-WP-TR-2012-0165** HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNATURE//
Raymond E. Slyh
Work Unit Manager
Human Trust and Interaction Branch

//SIGNATURE//
Louise A. Carter, PhD.
Human-Centered ISR Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

| | |
|---|---|
| **REPORT DOCUMENTATION PAGE** | *Form Approved*<br>**OMB No. 0704-0188** |

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports,
1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget,
Paperwork Reduction Project (0704-0188) Washington, DC 20503.
**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

| | | |
|---|---|---|
| **1. REPORT DATE** *(DD-MM-YYYY)*<br>29-11-2012 | **2. REPORT TYPE**<br>Interim | **3. DATES COVERED** *(From - To)*<br>1 October 2010 – 30 September 2012 |

| | |
|---|---|
| **4. TITLE AND SUBTITLE**<br>Foreign Language Analysis and Recognition (FLARe) Initial Progress | **5a. CONTRACT NUMBER**<br>FA8650-09-D-6939 |
| | **5b. GRANT NUMBER**<br>N/A |
| | **5c. PROGRAM ELEMENT NUMBER**<br>62202F |
| **6. AUTHOR(S)**<br>Brian Ore (SRA International Inc.)<br>Steve Thorn (SRA International Inc.)<br>Dave Hoeferlin (SRA International Inc.)<br>Raymond E. Slyh (711 HPW/RHXS)<br>Eric G. Hansen (711 HPW/RHXS) | **5d. PROJECT NUMBER**<br>5328 |
| | **5e. TASK NUMBER**<br>0028 |
| | **5f. WORK UNIT NUMBER**<br>H06K (5328X02S) |
| **7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**<br>Air Force Research Laboratory<br>711th Human Performance Wing, Human Effectiveness Directorate<br>2255 H Street<br>Wright-Patterson Air Force Base, OH 45433 | **8. PERFORMING ORGANIZATION REPORT NUMBER** |
| **9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**<br>Air Force Materiel Command<br>Air Force Research Laboratory<br>711th Human Performance Wing<br>Human Effectiveness Directorate<br>Human-Centered ISR Division<br>Wright-Patterson Air Force Base, OH 45433 | **10. SPONSOR/MONITOR'S ACRONYM(S)**<br>711 HPW/RHXS |
| | **11. SPONSORING/MONITORING AGENCY REPORT NUMBER**<br>**AFRL-RH-WP-TR-2012-0165** |

**12. DISTRIBUTION AVAILABILITY STATEMENT**
Approved for public release; distribution is unlimited.

**13. SUPPLEMENTARY NOTES**
88ABW-2013-0379, dated 29 January 2013

**14. ABSTRACT**
This interim report provides research results in the areas of automatic speech recognition (ASR), machine translation (MT), topic detection, natural language processing (NLP), and information retrieval (IR).

**15. SUBJECT TERMS**
Automatic speech recognition (ASR), machine translation (MT), topic detection, natural language processing (NLP), information retrieval (IR).

| | | | | | |
|---|---|---|---|---|---|
| **16. SECURITY CLASSIFICATION OF:**<br>Unclassified | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON**<br>Raymond E. Slyh |
| **a. REPORT**<br>U | **b. ABSTRACT**<br>U | **c. THIS PAGE**<br>U | SAR | 40 | **19b. TELEPONE NUMBER (*Include area code)***<br>N/A |

**THIS PAGE LEFT INTENTIONALLY BLANK.**

## TABLE OF CONTENTS

**Section** **Page**

## LIST OF FIGURES

## LIST OF TABLES

# SUMMARY

This document provides a summary of work completed by government researchers as well as SRA International under the work unit 5328X02S, Foreign Language Analysis and Recognition (FLARe). The government work was performed over the period of 1 October 2010 to 30 September 2012, while the SRA work was performed over the period 23 August 2011 to 30 September 2012 under contract FA8650-09-D-6939-0028.

Over this period, work was accomplished on Automatic Speech Recognition (ASR), Machine Translation (MT), and topic detection. Also, several modifications were made to an in-house Multilingual Multimedia Information Extraction and Retrieval (MMIER) system called Haystack.

For ASR, systems were developed on several languages and integrated into the Haystack system. Several experiments were conducted on methods to reduce the effects of Out-of-Vocabulary (OOV) words encountered by an ASR system, where OOV words are those words spoken by a person that are not in the pronunciation dictionary and Language Model (LM) for an ASR system. By definition, OOV words will not appear in the output of an ASR system, so they naturally increase Word Error Rates (WERs). This report also describes ASR systems built as part of the 2012 International Workshop on Spoken Language Translation (IWSLT) algorithm evaluation. For MT, the report describes experiments conducted in the course of developing Arabic-to-English MT systems for the 2012 IWSLT evaluation. For topic detection, the report describes several experiments conducted with detectors based on LMs and Support Vector Machines (SVMs). Finally, the report describes several modifications made to the processing pipeline, ASR systems, and user interface for the Haystack system.

# 1.0 INTRODUCTION

This document provides a summary of work completed by government researchers as well as SRA International under the work unit 5328X02S, FLARe. The government work was performed over the period of 1 October 2010 to 30 September 2012, while the SRA work was performed over the period 23 August 2011 to 30 September 2012 under contract FA8650-09-D-6939-0028.

Over this period, work was accomplished on ASR, MT, and topic detection. Also, several modifications were made to a MMIER system called Haystack, which has been developed in-house under prior efforts. For ASR, systems were developed on several languages and integrated into the Haystack system. Several experiments were conducted on methods to reduce the effects of OOV words encountered by an ASR system, where OOV words are those words spoken by a person that are not in the pronunciation dictionary and LM for an ASR system. By definition, OOV words will not appear in the output of an ASR system, so they naturally increase WERs. This report also describes ASR systems built as part of the 2012 IWSLT algorithm evaluation. For MT, the report describes experiments conducted in the course of developing Arabic-to-English MT systems for the 2012 IWSLT evaluation. For topic detection, the report describes several experiments conducted with detectors based on LMs and SVMs. Finally, the report describes several modifications made to the processing pipeline, ASR systems, and user interface for the Haystack system

This report is organized as follows. Section 2.0 describes experiments and accomplishments in ASR, MT, and topic detection. In addition, this section describes enhancements to the Haystack system. Section 3.0 summarizes conclusions drawn from the experiments and makes recommendations for future efforts.

## 2.0 EXPERIMENTS AND ACCOMPLISHMENTS

This section discusses several experiments and accomplishments for the covered period. Section 2.1 discusses a number of ASR experiments that were conducted. Several methods were examined to deal with the effects of OOV words on ASR WERs. This section also describes ASR systems that were built for the 2012 IWSLT algorithm evaluation. Section 2.2 discusses the development of an Arabic-to-English (AE) MT system developed for the IWSLT 2012 evaluation. Section 2.3 discusses experiments in topic detection. Section 2.4 discusses improvements made to the in-house Haystack MMIER system.

### 2.1 ASR Experiments

This section discusses how ASR systems were designed to reduce the effects of OOV words encountered by a recognizer. OOV words are those words spoken by a person that are not in the pronunciation dictionary and LM for an ASR system; as a result, they will never appear in the output of the recognizer, thereby increasing the WER. LMs were estimated using both words and sub-word units that can be combined to form words. Section 2.1.1 describes how graphones were used as sub-word units for English and Croatian ASR systems. Section 2.1.1 describes how morphemes were used as sub-word units for Arabic ASR systems. Section 2.1.2 describes how Morfessor and the Comprehensive Perl Archive Network (CPAN) Russian Stemmer were used to segment words for Russian ASR systems. Finally, Section 2.1.4 describes an ASR system developed for the ISWLT 2012 evaluation.

### 2.1.1. Graphones for English and Croatian ASR

Sub-word units for English and Croatian were derived from Grapheme-to-Phoneme (G2P) models estimated using the Sequitur G2P system [1, 2]. Sequitur models a word-pronunciation pair as a sequence of graphones $\boldsymbol{q} = \{q_1, q_2, \dots, q_N\}$, where $N$ is the number of graphones needed to represent a word-pronunciation pair, $q_n = (\boldsymbol{g}_n, \boldsymbol{p}_n)$ is the $n^{th}$ graphone, $\boldsymbol{g}_n$ is the $n^{th}$ grapheme (letter) sequence, and $\boldsymbol{p}_n$ is the $n^{th}$ phoneme sequence. Consider the following example for the word *caterpillar* with pronunciation /*k ae t axr p ih l axr*/. This word-pronunciation pair might be modeled by four graphones:

$$q_1 = ([c\ a], [k\ ae]),\quad q_2 = ([t\ e\ r\ p], [t\ axr\ p]),\quad q_3 = ([i\ l\ l], [ih\ l]),\quad q_4 = ([a\ r], [axr]).$$

Two parameters $L$ and $M$ are required when training the G2P models; $L$ is used to control the graphone size such that the number of graphemes in each $\boldsymbol{g}_n$ and phonemes in each $\boldsymbol{p}_n$ cannot exceed $L$, and $M$ is the $M$-gram order used to model sequences of graphones. Given a pronunciation dictionary and a text corpus that includes OOV words, the following procedure can be used to incorporate graphones into the recognizer:

- Train G2P models on the pronunciation dictionary for a specified *L* and *M*
- Add the graphones from the G2P models to the pronunciation dictionary
- Evaluate the G2P models on all OOV words from the text corpus
- Replace the OOV words in the text corpus with their graphone sequence
- Train an LM on the text that includes both words and graphones
- Evaluate the recognizer using the modified pronunciation dictionary and LM
- Merge the graphemes from consecutive graphones to form words

***English Results:*** The interaction between *L* and *M* was evaluated on English by training G2P models on 30K words from the Carnegie Mellon University (CMU) English pronunciation dictionary. These models were evaluated on a disjoint subset of the dictionary and the hypothesized pronunciations were compared to the reference pronunciations. Table 1 shows the Phoneme Error Rates (PERs) obtained. We can see that as ***L*** increases, the improvement obtained by using higher order *M*-grams diminishes. The best performance was obtained using $L = 1$ and $M = 5$.

Sub-word LMs were trained on the Wall Street Journal (WSJ) Continuous Speech Recognition (CSR) LM-1 corpus [3]. Three subsets of the CMU pronunciation dictionary were created using the top 5K, 20K, and 64K words from the LM corpus, and G2P models were developed using $M = 3$ and $L = 2, 3, 4, 5, 6$. The Stanford Research Institute Language Modeling (SRILM) Toolkit [4] was used to estimate trigram LMs for each subset and G2P model configuration.

Acoustic Models (AMs) were trained on the WSJ database [5, 6] using the Hidden Markov Model (HMM) ToolKit (HTK) [7]. All HTK systems discussed in this report modeled phonemes using state-clustered across-word triphone HMMs that were discriminatively trained using the Minimum Phone Error (MPE) criterion. The final HMM set included 4500 shared states with an average of 24 mixtures per state. The feature set consisted of 12 Mel-Frequency Cepstral Coefficients (MFCCs), plus energy, with mean normalization applied on a per utterance basis. Delta, acceleration, and third differential coefficients were appended to form a 52 dimensional feature vector, and Heteroscedastic Linear Discriminant Analysis (HLDA) was applied to reduce the feature dimension to 39.

Each set of models was evaluated on the ARPA 1993 and 1994 Hub-1 development data. Decoding was performed using the HTK Large Vocabulary Continuous Speech Recognizer (LVCSR) HDecode. Table 2 shows the results obtained. The baseline systems used LMs trained on words only. System performance was measured using WER and Letter Error Rate (LER). We can see that when using small vocabularies, the sub-word LM provides a substantial improvement in system performance. As the word vocabulary is increased, and the OOV rate is decreased, the sub-word LM provides less benefit.

Table 1: PERs obtained Using G2P Models on English

| *M* | *L*=1 | *L*=2 | *L*=3 | *L*=4 | *L*=5 | *L*=6 |
|---|---|---|---|---|---|---|
| 1 | 53.2 | 34.7 | 25.0 | 19.5 | 17.7 | 19.2 |
| 2 | 21.0 | 10.9 | 11.5 | 13.2 | 15.0 | 17.3 |
| 3 | 10.8 | 9.3 | 11.3 | 13.2 | 15.0 | 17.3 |
| 4 | 8.3 | 9.3 | 11.3 | 13.2 | | |
| 5 | 8.0 | 9.3 | 11.3 | | | |
| 6 | 8.0 | | | | | |

Table 2: English Results Using Graphone Sub-Word LMs

| | Vocabulary | | 1993 and 1994 Hub-1 Development | | |
|---|---|---|---|---|---|
| Model | Words | Graphones | OOV | WER | LER |
| Baseline 5K | 5002 | | 1520 | 24.2 | 10.5 |
| 5K, L=2 | 5K+ | 1021 | (9.7%) | 16.5 | 7.3 |
| 5K, L=3 | 5K+ | 2804 | | 14.8 | 6.4 |
| 5K, L=4 | 5K+ | 3972 | | 16.0 | 6.9 |
| 5K, L=5 | 5K+ | 4156 | | 17.2 | 7.6 |
| 5K, L=6 | 5K+ | 4118 | | 18.0 | 8.0 |
| Baseline 20K | 19996 | | 391 | 12.4 | 5.4 |
| 20K, L=2 | 20K+ | 1297 | (2.5%) | 11.8 | 5.5 |
| 20K, L=3 | 20K+ | 5772 | | 10.4 | 4.6 |
| 20K, L=4 | 20K+ | 10796 | | 10.4 | 4.6 |
| 20K, L=5 | 20K+ | 12957 | | 10.5 | 4.6 |
| 20K, L=6 | 20K+ | 13462 | | 10.8 | 4.7 |
| Baseline 64K | 63982 | | 92 | 9.9 | 4.4 |
| 64K, L=2 | 64K+ | 1567 | (0.6%) | 10.2 | 4.4 |
| 64K, L=3 | 64K+ | 9926 | | 9.8 | 4.4 |
| 64K, L=4 | 64K+ | 24136 | | 9.7 | 4.3 |

***Croatian Results:*** The experiments discussed in the "English Results" section on page 4 were repeated on Croatian using the GlobalPhone corpus and pronunciation dictionary [8]. The interaction between *L* and *M* was evaluated by training models on 17K words from the GlobalPhone pronunciation dictionary. Table 3 shows the PERs obtained. Compared to the results in Table 1, we can see that lower PERs were obtained for Croatian than English, and increasing *M* yields more improvement for English than for Croatian. This result can be attributed to the fact that Croatian has a more phonemic orthography than English. The best performance was obtained using *L* = 1 and *M* = 2.

Two subsets of the GlobalPhone Croatian pronunciation dictionary were created using the top 5K and 10K words from the GlobalPhone corpus. Note that it was not possible to use the same vocabulary sizes as in the English experiments, because the LM corpus only includes 18K unique words. G2P models and trigram LMs were developed for each subset using the same procedure described in the "English Results" section on page 4.

The AMs were trained using HTK. The final HMM set included 1500 shared states with an average of 16 mixtures per state. The feature set consisted of 12 Perceptual Linear Prediction

(PLP) coefficients, plus the zeroth coefficient, with mean normalization applied on a per utterance basis. Delta and acceleration coefficients were appended to form a 39 dimensional feature vector.

Table 4 shows the results obtained on the GlobalPhone development partition. As was the case with English, the sub-word LM provides a substantial improvement when using small vocabularies. However, the sub-word systems do not outperform the baseline 18K system that was created by simply training an LM on all of the available words. Compared to the results in Table 2, we can see that the OOV rates are higher in the GlobalPhone Croatian corpus than the English WSJ corpus. This result suggests that it would be useful to use a larger text corpus.

**Table 3: PERs Obtained Using G2P Models on Croatian**

| *M* | *L=1* | *L=2* | *L=3* | *L=4* | *L=5* | *L=6* |
|---|---|---|---|---|---|---|
| 1 | 4.9 | 3.6 | 3.6 | 4.5 | 6.8 | 9.8 |
| 2 | 3.0 | 4.1 | 4.5 | 5.4 | 8.2 | 10.5 |
| 3 | 3.5 | 3.7 | 4.4 | 5.4 | 8.2 | 10.5 |
| 4 | 3.5 | 3.6 | 4.5 | 5.4 | | |
| 5 | 3.5 | 3.6 | 4.5 | | | |
| 6 | 3.2 | | | | | |

**Table 4: Croatian Results Using Graphone Sub-Word LMs**

| | **Vocabulary** | | **GlobalPhone Development** | | |
|---|---|---|---|---|---|
| **Model** | **Words** | **Graphones** | **OOV** | **WER** | **LER** |
| Baseline 5K | 5002 | | 3715 | 54.0 | 18.4 |
| 5K, L=2 | 5K+ | 463 | (25.1%) | 46.0 | 14.3 |
| 5K, L=3 | 5K+ | 2138 | | 45.2 | 13.9 |
| 5K, L=4 | 5K+ | 5327 | | 44.8 | 13.8 |
| 5K, L=5 | 5K+ | 7658 | | 44.6 | 13.6 |
| 5K, L=6 | 5K+ | 8567 | | 44.2 | 13.7 |
| Baseline 10K | 10002 | | 2902 | 46.3 | 15.4 |
| 10K, L=2 | 10K+ | 402 | (19.6%) | 44.0 | 13.7 |
| 10K, L=3 | 10K+ | 1910 | | 43.8 | 13.5 |
| 10K, L=4 | 10K+ | 4381 | | 43.4 | 13.4 |
| 10K, L=5 | 10K+ | 5910 | | 42.9 | 13.3 |
| 10K, L=6 | 10K+ | 6334 | | 42.7 | 13.4 |
| Baseline 18K | 17772 | | 2234<br>(15.1%) | 41.4 | 13.6 |

### 2.1.1. Morphemes for Arabic ASR

Arabic is a morphologically rich language and Arabic speech recognizers typically need much larger vocabularies than their English counterparts to have similar OOV rates. As a result, it is more difficult to estimate robust Arabic LMs, because the high number of possible word forms leads to sparse training data. Sub-word LMs can help to alleviate these problems by modeling morphemes instead of words.

**Table 5: Arabic WERs Using Sub-Word LMs**

| System Description | WER |
|---|---|
| Full morphological decomposition | 25.4 |
| Merged consecutive prefixes | 25.4 |
| No decomposition words with stems shorter than 3 characters | 25.1 |
| No decomposition for the most frequent 5K decomposable words | 24.6 |
| No decomposition for the most frequent 10K decomposable words | 24.6 |

The MADA tool [9] was used to perform morphological analysis and disambiguation on text from the Global Autonomous Language Exploitation (GALE)[1] and Topic Detection and Tracking (TDT) [10, 11] corpora. MADA outputs the vowelized form of words and the decomposition of words into morphemes. Note that the vowelization process makes it possible to create a pronunciation dictionary from the words using a set of rules. The following procedure was used to create a morpheme-based system:

- Evaluate MADA on the text corpus
- Create a pronunciation dictionary from the vowelized morphemes
- Train an LM on the unvowelized morphemes, where a "+" sign is attached to the end of prefixes and to the start of suffixes, so that they can be reattached after running the ASR system
- Evaluate the recognizer using the pronunciation dictionary and sub-word LM
- Reattach prefixes and suffixes

Additional systems were also created by imposing the following rules to the morphological decomposition [12]: (1) consecutive prefixes were merged into a single token; (2) words with stems shorter than three characters were not decomposed into morphemes; and (3) the most frequent *N* words were not decomposed into morphemes.

The AMs were trained on 251 hours from the GALE corpus using HTK. The final HMM set included 6000 shared states with an average of 24 mixtures per state. The feature set consisted of 12 PLPs, plus the zeroth coefficient, with mean normalization applied on a per utterance basis. Delta, acceleration, and third differential coefficients were appended to form a 52 dimensional feature vector, and HLDA was applied to reduce the feature dimension to 39.

Each set of models was evaluated on the GALE Phase II development partition. HDecode was evaluated using a trigram LM and the recognition lattices were rescored with a 4-gram LM. Table 5 shows the WERs obtained. We can see that an improvement in system performance was obtained by not decomposing words with short stems and not decomposing the most frequent 5K or 10K words.

### 2.1.2. Morfessor and CPAN Russian Stemmer for Russian ASR

Russian speech recognizers typically need very large vocabularies to have low OOV rates. This section investigates two programs for automatically deriving Russian sub-word units. Morfessor [13] segments words into morpheme-like units that are automatically learned from a text

[1] http://projects.ldc.upenn.edu/gale/

**Table 6: Russian WERs Using Sub-Word LMs**
**Dashes indicate that CoMMA was not applied**

| Sub-Word Analysis | CoMMA Threshold | WER |
|---|---|---|
| None | - | 26.2 |
| Morfessor | 20 | 25.7 |
| | 100 | 25.8 |
| | 1000 | 26.3 |
| | - | 27.6 |
| CPAN Russian Stemmer | 20 | 25.7 |
| | 100 | 25.7 |
| | 1000 | 26.3 |
| | - | 28.2 |

database, and the CPAN Russian Stemmer[2] applies the Porter stemming algorithm to derive the root of a word. The following procedure was used to incorporate these sub-word units into the recognizer:

- Evaluate Morfessor or the CPAN Russian Stemmer on the text corpus
- Create a pronunciation dictionary by evaluating a G2P model on the sub-word units
- Train an LM on the sub-word units, and attach a "+" sign to the start of the every sub-word unit except for the first sub-word unit from a word
- Evaluate the recognizer using the pronunciation dictionary and sub-word LM
- Attach sub-word units that start with a "+" sign to the previous word or sub-word unit

This procedure was applied to text from the GlobalPhone corpus, the Open Subtitles corpus [14], and articles downloaded from Wikipedia.[3] The G2P model was trained on the GlobalPhone Russian pronunciation dictionary using the Sequitur G2P system. Additional systems were also created by applying Count Mediated Morphological Analysis (CoMMA) [15]. CoMMA prevents a word from being segmented into sub-word units if the word appears more than *N* times in the training text, where *N* is a threshold chosen by the user. Larger values for *N* result in more words being segmented. All systems used a 400K vocabulary.

The AMs were trained on the GlobalPhone corpus using HTK. The final HMM set included 1500 shared states with an average of 16 mixtures per state, and the PLP features were calculated using the same procedure described in Section 2.1.1 A second set of models was estimated that include Speaker Adaptive Training (SAT).

Each set of models was evaluated on the GlobalPhone development partition. Initial transcripts were produced using HDecode with the non-SAT HMMs. Constrained Maximum Likelihood Linear Regression (CMLLR) transforms were estimated for each speaker, and recognition lattices were generated using the SAT HMMs. The final transcripts were created by rescoring the lattices with a 4-gram LM. Table 6 shows the WERs obtained. Compared to the baseline

[2] Available at: http://search.cpan.org/~algdr/Lingua-Stem-Ru-0.01/Ru.pm
[3] Available at: http://dumps.wikimedia.org/ruwiki/

system trained on only words, the sub-word LMs yielded an improvement in WER when CoMMA is applied with $N = 20$ and $N = 100$.

### 2.1.3. IWSLT 2012 ASR System

This section discusses the ASR systems that were developed for the IWSLT 2012 evaluation. One set of AMs was trained using PLP features, and a second set of AMs was trained using MFCC features. The acoustic data consisted of 807 TED Talks that were downloaded from the internet [30], and the HMMs were estimated using HTK. The final HMM sets included 6000 shared states with an average of 28 mixtures per state. The feature set included 12 MFCCs or PLPs, plus the zeroth coefficient, with mean normalization applied on a per utterance basis. Delta, acceleration, and third differential coefficients were appended to form a 52 dimensional feature vector, and HLDA was applied to reduce the feature dimension to 39. A second set of HMMs was trained for each feature set that included SAT.

LMs were developed on the English Gigaword corpus [31] and the following data sets provided by IWLST: Europarl, news 2007–2011, news commentary, and TED. Cross-entropy difference scoring [32] was used to select subsets of the text for LM training. Given an in-domain text corpus and a larger out-of-domain text corpus, the following procedure can be used to apply cross-entropy difference scoring. First, estimate an in-domain LM on the in-domain text corpus. Denote the cross-entropy of a sentence $s$ according to this LM as $H_I(s)$. Next, randomly sample a subset of the out-of-domain text corpus that is similar in size the in-domain text corpus. Estimate an LM on this subset and denote the cross-entropy of a sentence $s$ according to this LM as $H_O(s)$. Finally, score each sentence $s$ from the out-of-domain text corpus according to $H_I(s) - H_O(s)$. These scores can be used to select a subset of the corpus by only using the top scoring sentences.

In the context of IWSLT, the TED data is in-domain and all other corpora are out-of-domain. Cross-entropy difference scoring was used to score each sentence from the Gigaword, Europarl, news 2007–2011, and news commentary data sets. Subsets were created for each out-of-domain corpus by selecting a fraction $N$ of the sentences. Trigram LMs were estimated for each subset, the perplexity was evaluated on the IWSLT dev2010 partition, and the value of $N$ that yielded the lowest perplexity was chosen as the selection threshold. This process selected 7.3% of the out-of-domain data for LM development.

Interpolated trigram and 4-gram LMs were estimated on the selected data using the SRILM Toolkit. Recurrent Neural Network Maximum Entropy (RNNME) LMs [33] were also developed using the RNNLM Toolkit.[4] One RNNME LM was trained on the selected Gigaword data, and a second RNNME LM was trained on the selected News 2007–2011 data. Each network included 160 hidden units, 300 classes, and $10^9$ direct connections. The LM vocabulary included 95K words.

Recognition lattices were generated using the same procedure as used in Section 2.1.2, and 1000-best lists were extracted for rescoring with the 4-gram and RNNME LMs. Prior to scoring the 1000-best lists, repeated words were merged into a single token. The scores from each LM were linearly interpolated using weights chosen to minimize the perplexity on the IWSLT development partitions. Table 7 shows the WERs obtained for the MFCC and PLP systems on the IWSLT dev2010 and tst2010 partitions. For comparison purposes, the WERs obtained by

[4] Available at: http://www.fit.vutbr.cz/~imikolov/rnnlm/

rescoring the 1000-best lists with only the 4-gram LM are included. We can see that rescoring the recognition lattices yielded a substantial improvement in system performance, and that the MFCC and PLP systems perform comparably.

**Table 7: WERs Obtained on the IWSLT dev2010 and tst2010 Partitions Using the MFCC and PLP ASR Systems**

| System | dev2010 | | tst2010 | |
|---|---|---|---|---|
| | MFCC | PLP | MFCC | PLP |
| First Pass | 19.0 | 18.3 | 18.7 | 17.9 |
| Second Pass | 16.6 | 16.5 | 15.4 | 15.0 |
| 4-gram LM Rescore | 15.3 | 15.4 | 14.1 | 13.9 |
| 4-gram LM + RNNME Rescore | 14.4 | 14.4 | 13.0 | 12.5 |

## 2.2 IWSLT 2012 AE MT System

This section describes discusses the development of some AE MT systems for the IWSLT 2012 evaluation. In previous AE MT systems for prior year IWSLT evaluations [15, 30, 34–36], we normalized various forms of Arabic alef and hamza and removed the Arabic tatweel character and some diacritics before applying a light Arabic morphological analysis procedure that we called AP5. For the 2012 evaluation, we modified the AP5 procedure to more closely conform to the Arabic Treebank (ATB) segmentation format used in the MADA Arabic morphological analysis, diacritization, and lemmatization system [37]. In [38], it was shown that the ATB format performed the best of the various MADA segmentation formats tried on the IWSLT 2011 evaluation. In particular, we kept the definite article (Al-) attached to its corresponding noun or adjective. We denote this modified AP5 system as AP5ATBLite.

Table 8 shows the mean Bilingual Evaluation Understudy (BLEU) scores for individual AE MT systems trained on the 2011 and 2012 training data and tested on the tst2010 data versus the morphology segmentation system. For both the 2011 and 2012 training data, the AP5ATBLite system performs slightly better than the AP5 system. Also, the extra training data in the 2012 system provides approximately one BLEU point of improvement over the systems trained on the 2011 data.

In addition to the AP5ATBLite modification, we investigated the use of Kneser-Ney (KN) phrase table smoothing [39] using the AP5ATBLite system trained on the 2012 training data. The combination of AP5ATBLite and KN smoothing yielded a mean BLEU score of 23.60 compared to the mean of 22.45 for the AP5ATBLite system without phrase table smoothing. The overall submitted AE system was a combination of individual component systems that were each the best in terms of BLEU score after ten Minimum Error Rate Training (MERT) optimization runs. Two of the component systems were (1) the best AP5ATBLite system (with no phrase table smoothing) and (2) the best AP5ATBLite System with KN phrase table smoothing.

**Table 8: Mean BLEU Scores for Individual AE MT Systems Tested on the tst2010 Data Versus Morphology Segmentation System and Year of Training Data)**

| Morphology System | Training Data Year | |
|---|---|---|
| | 2011 | 2012 |
| AP5 | 21.13 | 22.24 |
| AP5ATBLite | 21.57 | 22.45 |

## 2.3 Topic Detection Experiments

Topic detection is defined here as the automatic categorization of an unknown text document. Topic categories can be broad (business, sports, science, political) or event specific (Hurricane Katrina, 2012 Summer Olympics, *etc*.). The focus of the following experiments is to automatically categorize documents into broad topics. The purpose of this categorization is twofold: (1) affiliating a topic category with specific text in the Haystack system can further enhance search and retrieval and (2) properly categorizing text can support identifying specific domains and guide the use of domain-specific LMs for both ASR and statistical MT systems. Current ASR and statistical MT systems perform their best when they have a narrow domain to work with, so if a topic detection system could confidently identify the domain of a set of text, then a second, domain-specific ASR or statistical MT pass could be run that would (hopefully) increase the final accuracy of the system.

A challenge for the task of topic detection is that a document often does not contain just a single topic. A document that discusses the performance of Wall Street during presidential election years can fit in the category of business and politics. Having to make a hard decision on identifying a particular document's topic for system training or testing purposes is not a straightforward task.

### 2.3.1. Data

Documents were gathered from all available TDT data sets [16]. Documents are available in the following languages: English (ENG), Arabic (ARB), and Mandarin Chinese (MAN). The TDT data sets were designed to evaluate the task of identifying and tracking specific events like Hurricane Katrina or the 2012 Summer Olympics. However, there are also general topic categories for each document as shown in Table 9.

### 2.3.2. Text Normalization

Preliminary text normalization and tokenization was accomplished before any experimentation. Various levels of tokenization were experimented with, and the final version used for all the experiments and results discussed in this document can be summarized in the following steps:

- HTML and XML tags removed
- General punctuation and special characters are removed, but ".", "!", and "?" are retained
- Compound words (*i.e.,* those words containing "-", "/", or "_") are split apart
- Possessives are removed
- Words beginning with "$" are replaced with "CUR*"
- Words ending in "%" are replaced with "PER*"
- Any remaining numbers are replaced with "NUM*," including words like 9mm, 401k, 10th, *etc.*

After tokenization, the PERL implementation of the Porter Stemmer [17] was applied. According to Dr. Martin Porter, the author of the Porter Stemmer algorithm, the definition of the stemmer is, "a process for removing the commoner morphological and inflectional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems." In general, the stemmer removes plurals and common suffixes and attempts to collapse words to their roots.

After formatting, the complete vocabulary covering all topics consisted of 64,218 unique words.

### 2.3.3. Classifiers and Feature Sets

Three different classifiers were considered in these experiments: (1) the CMU Language Modeling Toolkit (CMU-LMTK) [18], (2), Torch [19], and (3) LIBSVM [20]. The CMU-

**Table 9: General Topic Categories**

| Number of Documents | Short Topic Designator | Full Topic Name |
|---|---|---|
| 1418 | ac | Accidents |
| 2143 | vw | Acts of Violence or War |
| 1190 | ch | Celebrity Human Interest News |
| 3254 | en | Elections |
| 1459 | fn | Financial News |
| 4875 | lc | Legal/Criminal Cases |
| 7436 | mn | Miscellaneous News |
| 1198 | nd | Natural Disasters |
| 2371 | nl | New Laws |
| 1716 | pd | Political and Diplomatic Meetings |
| 1014 | sh | Scandals Hearings |
| 1008 | sd | Science and Discovery News |
| 1786 | sp | Sports News |

LMTK was used because of familiarity with it from previous work on speaker recognition. Torch and LIBSVM were used as both are SVM toolkits, and SVMs are very often used for text categorization due to their very good performance.

The CMU/LMTK-based classifiers were derived from statistical LMs built from the various topic categories. First, LMs were built for each topic category by merging all documents for a given category and building a unigram language model with Witten-Bell smoothing. Second, a background LM was built by combining all of the documents (across all of the topics) and building a unigram language model, again with Witten-Bell smoothing. When testing a document for its topic, we computed the perplexity of the document against each topic model normalized by the perplexity of the background model. Log probabilities that exist in both a given topic model and the background model are summed. The final document score for a given topic is determined by the sum of log probabilities normalized by the number of valid (existing) probabilities. The topic with the highest score is chosen as the topic for the document.

For the SVM-based systems, each document was represented as a single feature vector for a given experiment, and there were two main types of document vectors considered. The first type of document vector consisted of normalized term frequencies (denoted as "tf"). Let $\{w_1, w_2, \dots, w_T\}$ be the set of all terms across all of the $D$ documents that will be considered for training, where $T$ is the total number of terms to be considered. For term $w_t$, its normalized tf value for document $d$ is given as

$$\text{Norm-tf}(w_t, d) = C(w_t, d) \Big/ \sum_{t=1}^{T} C(w_t, d)$$

where $C(w_t, d)$ is the number of occurrences of term $w_t$ in document $d$. The second type of document vector consisted of normalized term frequencies times inverse document frequencies (denoted as "tf-idf"), where the inverse document frequencies are first scaled by a log function. Thus, the normalized tf-idf value for term $w_t$ in document $d$ is given as

$$\text{Norm-tf-idf}(w_t, d) = \text{Norm-tf}(w_t, d) * \log_{10}\left(D / \hat{C}(w_t)\right)$$

where $\hat{C}(w_t)$ is the number of *documents* in the training collection that contain the term $w_t$. For the normalized tf, all terms are considered equally important. This feature generally has less discriminating power to determine the relevance of terms across documents compared to the tf-idf feature, which is why tf-idf is the most common feature type for text categorization methods. According to [21], the properties of tf-idf include: (1) a value that is highest when the term occurs many times within a small number of documents, (2) a value that is lower when the term occurs fewer times in a document (or occurs in many documents), and (3) a value that is lowest when the term occurs in virtually all documents.

For the SVM-based systems, every document is represented as a vector. To create this vector, a set of vocabulary terms must first be defined. As previously mentioned, after text normalization, 64,218 unique words defined the entire data set. Each word was assigned a position in the vector and the value at that position was either the normalized tf or tf-idf value as previously defined. Note that this process does lock the vocabulary during training. If a document is evaluated that contains words not included during training, these words are ignored and do not contribute to the final topic detection score. Expanding the vocabulary requires retraining of all the models.

For training an SVM model for a given topic, each document vector is assigned either a "+1" for a "true" topic case (*i.e.*, the document belongs to the topic for which the model is being trained) or a "-1" for a "false" topic case (*i.e.*, the document does not belong to the topic for which the model is being trained). This set of training vectors and its corresponding set of "+1" and "-1" output labels are then used to train an SVM with a linear kernel (specifically, a linear kernel epsilon based regression model for LIBSVM) for each topic category. Note that a given document vector will be assigned an output label of "+1" when that document is used as a positive example of its proper topic and "-1" when that document is used as a negative example for the other topics.

To evaluate a document with SVM topic models, it is first converted into its document vector format and then evaluated against each topic SVM model. The resulting scores from the topic models are ordered to determine the best topic match.

### 2.3.4. Experiments

All experiments used a Fisher-Yates randomization of the files and divided the file lists into 80% for training and 20% for testing. Each "experiment" and the documented classification error are the result of segmenting the data this way 100 times and then aggregating the results.

Initial experiments used all 13 topic categories. Preliminary results showed a large amount of confusion with the "Miscellaneous News (mn)" category. This category contained a large number of documents, and it was a very diverse collection of topics. This category was removed from the data set, and the topic detection performance improved.

Later, during development of the fully trained topic models for integration with the Haystack system, it was found that "Legal/Criminal Cases (lc)" exhibited behavior similar to that of the mn category. The lc category also had a lot of documents, and the document content tended to be vague. This topic category was also removed during later experimentation.

The Equal Error Rates (EERs) (*i.e.*, the error rates found when thresholds are adjusted such that the probability of a miss is equal to the probability of a false alarm) listed in this section do not include any effects from the mn category. Results for cases where the lc category was removed are labeled as "no lc".

Figure 1 is a summary of aggregated EERs across each topic category for five different topic detection systems: (1) CMU-LMTK unigram LM, (2) Torch SVM with tf features, (3) Torch SVM with tf-idf features, (4) Torch SVM with tf-idf features and "no lc," and (5) LIBSVM with tf-idf features and "no lc." As seen in Figure 1, different topics exhibit different performance ranges. "Financial News," "Sports," and "Science and Discovery" topics consistently rank in the best performing topic categories across the different classification systems. These topics tend to be well defined and do not have much overlap with the other ten categories. The worst performing topic categories include "Violence and War," "Election News," "Political Discussions," and "Legal/Criminal Cases." Documents in these categories tend to overlap each other often with respect to vocabulary and can prove difficult to make a hard decision on a single topic. "Celebrities and Human Interest" is also a poor performing topic category; a potential reason for this could be that this is a poorly defined topic. Typical documents in this category contain names that fit under the term "celebrity," but these names are discussed in relation to specific events–usually scandals or political stories. Also, from confusion matrices of these results, one finds that the topic categories “Accidents” and “Natural Disasters” are often confused.

Regarding the different classification systems, it can be seen that with the exception of “Violence and War” and “Election News,” SVM systems outperform the unigram LM systems. Moving from the CMU-LMTK system to a basic normalized term frequency (tf) SVM model using Torch provides an EER reduction in the best performing eight topics. Using the normalized term frequency with inverse document frequency (tf-idf) features improves performance across every topic category. Lastly removing the lc data from the experiments either displayed a performance improvement or no change to every topic category. For comparison purposes, a LIBSVM (tf-idf with “no lc”) system was evaluated, and it performed on par or better, depending on the topic, than the Torch SVM system. For the end application of integration with the Haystack system, the LIBSVM system was the preferred system due to its licensing terms.

To further analyze the performance of the LIBSVM system, each topic category's performance was plotted on a Detection Error Trade-off (DET) plot as seen in Figure 2. Depending on where a system is expected to operate along the trade-off between false alarms vs. misses, it can be seen that EER does not tell the whole story. In general, these curves do show consistent behavior and gave us the confidence to move forward with deploying the topic detectors in the Haystack system.

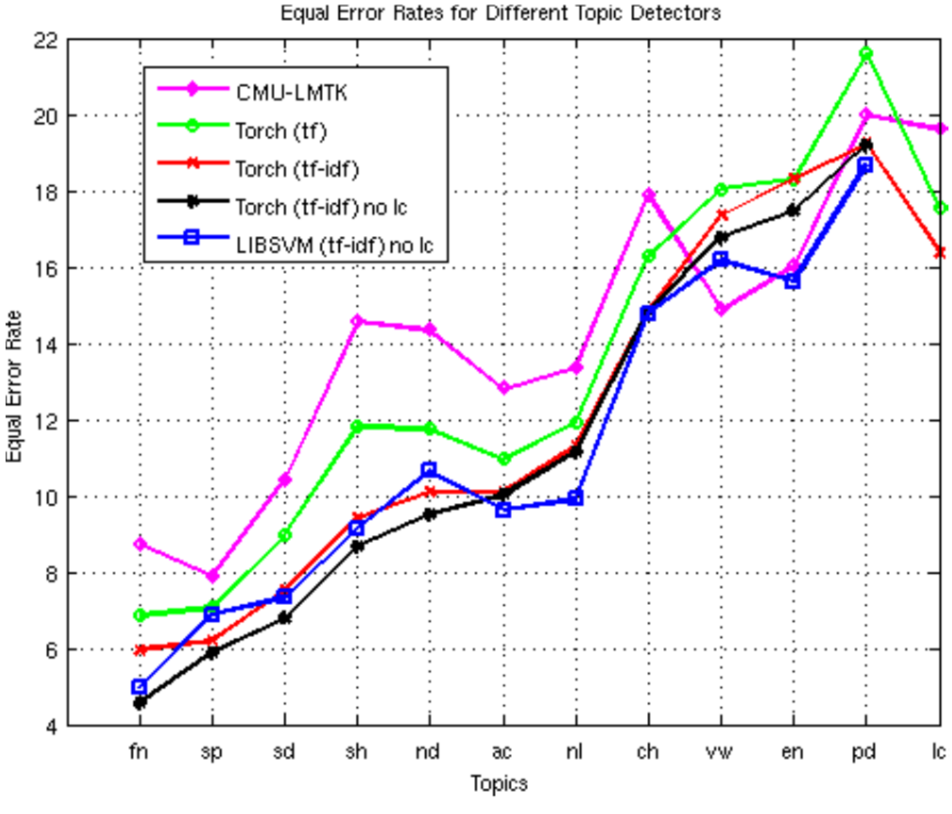
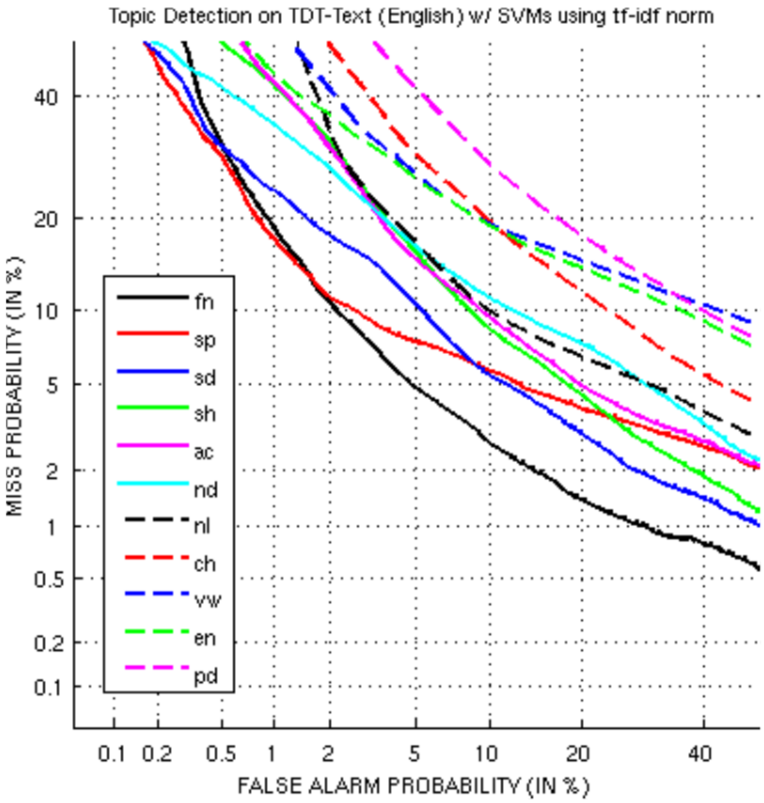


**Figure 1: Summary of Aggregated EER across each Topic Category and for Five Different Topic Detection Systems**



System

**Figure 2: DET Plot for the LIBSVM (tf-idf with “no lc”) System Showing the Performance of the Per Topic Category; Topic Classes are Ordered from Best to Worst with Respect to EER**

## 2.4 Haystack MMIER System

This section describes improvements made to the in-house Haystack MMIER system initially developed under a prior work unit.

### 2.4.1. Pipeline Improvements

This section discusses the improvements that were made to the Haystack pipeline. First, the ASR pipeline was modified to support different ASR systems for the first and second pass recognizer. This modification allows the use of non-SAT HMMs in the first pass and SAT HMMs in the second pass. Another benefit is that smaller LMs can be used in the first pass to speed up the recognition process. Functionality was also added to the ASR pipeline to support cascading global and class-based feature transforms.

Janya’s Semantex system was integrated into the pipeline for performing English and Chinese Named Entity (NE) extraction. NE extraction is performed on text documents and ASR transcripts when the source language is English or Chinese. For languages other than English, NE extraction is performed on the English translation generated by Language Weaver, Systran, and/or Moses-based MT systems.

The speaker diarization code was optimized to execute faster and yield a lower Diarization Error Rate (DER). Minimizing the file read and write operations reduced the execution time on a 30-

minute broadcast from 1 minute to 7 seconds. Optimizing the speaker segmentation and clustering thresholds across several files reduced the overall DER from 27.4% to 19.5%.

### 2.4.2. ASR Systems

This section discusses the Farsi, French, German, Portuguese, Russian, Spanish, and Turkish ASR systems that were developed for Haystack. The AMs were trained on the following corpora: Farsi from the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) corpus; French, German, Portuguese, Russian, and Spanish from the GlobalPhone corpus; and Turkish from the Turkish Broadcast News corpus [22]. All systems were trained using HTK, and the final HMM sets included 1500–4500 shared states with an average of 16–24 mixtures per state. A second set of HMMs was estimated for each language that included SAT. The Farsi, French, German, Portuguese, Russian, and Turkish models used the PLP feature set described in Section 2.1.1, and the Spanish models used the MFCC feature set described in Section 2.1.1. An additional set of narrowband models was estimated for Farsi, French, German, Portuguese, Russian, and Turkish by limiting the filter bank analysis to 125–3800 Hz.

The SRILM Toolkit was used to estimated trigram and 4-gram LMs for each language. In addition to the transcripts used for training the AMs, LMs were estimated using the following data sources: Farsi from the TRANSTAC corpus and Tehran English-Persian (TEP) Parallel corpus [23]; French from the French Gigaword corpus[24]; German from the European Language Newspaper Text corpus [25] and Wikipedia;[5] Portuguese from the CETEMPúblico corpus [26], the Portuguese Newswire Text corpus [27], and Wikipedia;[6] Russian from the Open Subtitles corpus and Wikipedia; Spanish from the Spanish Gigaword corpus [28] and the Spanish Broadcast News Transcripts corpus [29]; and Turkish from Wikipedia.[7]

[5] Available at: http://dumps.wikimedia.org/dewiki/

[6] Available at: http://dumps.wikimedia.org/ptwiki/

[7] Available at: http://dumps.wikimedia.org/trwiki/

**Table 10: WERs Obtained for Each Language; LM1 was Estimated from the AM Transcripts, and LM2 was Estimated from the AM Transcripts plus Additional Data**

| System | LM Tokens | OOV Rate | WER |
|---|---|---|---|
| Farsi LM1 | 146k | 4.4 | 31.2 |
| Farsi LM2 | 4M | 1.7 | 28.7 |
| French LM1 | 224k | 3.8 | 22.1 |
| French LM2 | 675M | 0.7 | 11.3 |
| German LM1 | 118k | 12.3 | 33.6 |
| German LM2 | 505M | 2.1 | 11.1 |
| Portuguese LM1 | 178k | 11.4 | 34.7 |
| Portuguese LM2 | 332M | 1.0 | 17.3 |
| Russian LM1 | 142k | 13.3 | 33.7 |
| Russian LM2 | 216M | 2.8 | 18.7 |
| Spanish LM1 | 146k | 8.8 | 27.9 |
| Spanish LM2 | 816M | 1.0 | 11.2 |
| Turkish LM1 | 737k | 4.9 | 19.2 |
| Turkish LM2 | 35M | 2.7 | 16.7 |

Decoding was performed using the same procedure described in Section 2.1.2. Table 10 shows the WERs obtained for each language. We can see that using additional text data to train the LMs yielded a substantial improvement in system performance for all languages.

### 2.4.3. Topic Detection for the Haystack System

The topic detection experiments described in Section 2.3 used documents from the TDT database. These documents were "well formed" or at least one could say they were "complete." The Haystack system is used to process broadcast news recordings and general video and audio files in addition to various types of text documents. The files that contain audio tracks are first processed by an ASR system to provide a transcript which then can be analyzed for topic detection. The output of the ASR, which would be the input to the topic detector for media files in the same language as the topic detector, is very different from the types of input that the topic detection systems of Section 2.3 were trained on. There are not necessarily any defined borders or story segments in the ASR transcripts, and the ASR transcripts generally contain errors. To overcome these problems, a sliding window is projected on the ASR output and the test vector for the LIBSVM models is computed on this sliding window. Window size and overlap ranges were evaluated, and it was determined that a two-minute window with a 30-second overlap provided usable results (consecutive windows with topic labels that are the same are merged for display purposes).

When the existing Haystack entries were processed through the LIBSVM topic detector, the following issues were found:

- The TDT categories do not match well with "real news."
- Models trained on complete documents do not match with windowed story segments as the test vectors are very sparse. Note that expanding the sliding window can remedy the data sparseness, but then the system is very slow to react to changing story lines that can happen frequently in typical news content.
- TDT categories are missing some obvious broad, but meaningful, categories (*e.g.,* medical, weather, religion, and technology).

In light of these issues, an entirely new topic detection paradigm was built for the Haystack system. First, new topic categories were added. Second, the entire SVM modeling framework was replaced with keyword spotting/counting. Third, the topic detector was built to accept Haystack-formatted text by default (with start-time/end-time/text-content). Finally, the windowing scheme was built into the topic detector right from the beginning. All of the text formatting steps and lessons learned from the prior experimentation were retained. The current nine topic categories implemented in the topic detection running in Haystack are:

- Business
- Entertainment
- Medical
- Political
- Religion
- Sports
- Technology
- Weather
- War/Military
- None of the Above

These topics are defined by a unique list of 40 keywords per topic.

The topic detection process now proceeds as follows:

- Process the ASR generated text (Format text – Porter Stemming – Compute word frequencies)
- Group text into windows (currently, two minutes with 30 second overlap)
- Match the input windowed word counts to "topic models"
- Merge consecutive topic labels that are the same
- Output final labels with start/end tags in a Haystack compatible format

Processing the Haystack database with this method yielded topic labels that matched very well with the data. This method, while simple, is extremely fast and allows for considerable future flexibility (*e.g.,* adding words to the keyword lists is much simpler then modifying the baseline vocabulary for the SVM-based systems and having to retrain all the models). Future work entails adding to each topic's keyword list to try to increase robustness. Also, this system was only developed for English; therefore, all foreign text must be translated before topic detection can be utilized. Porting these keyword lists to other languages so as to run the topic detector prior to translation could help for the foreign language media files, and this process would be relatively straightforward to do.

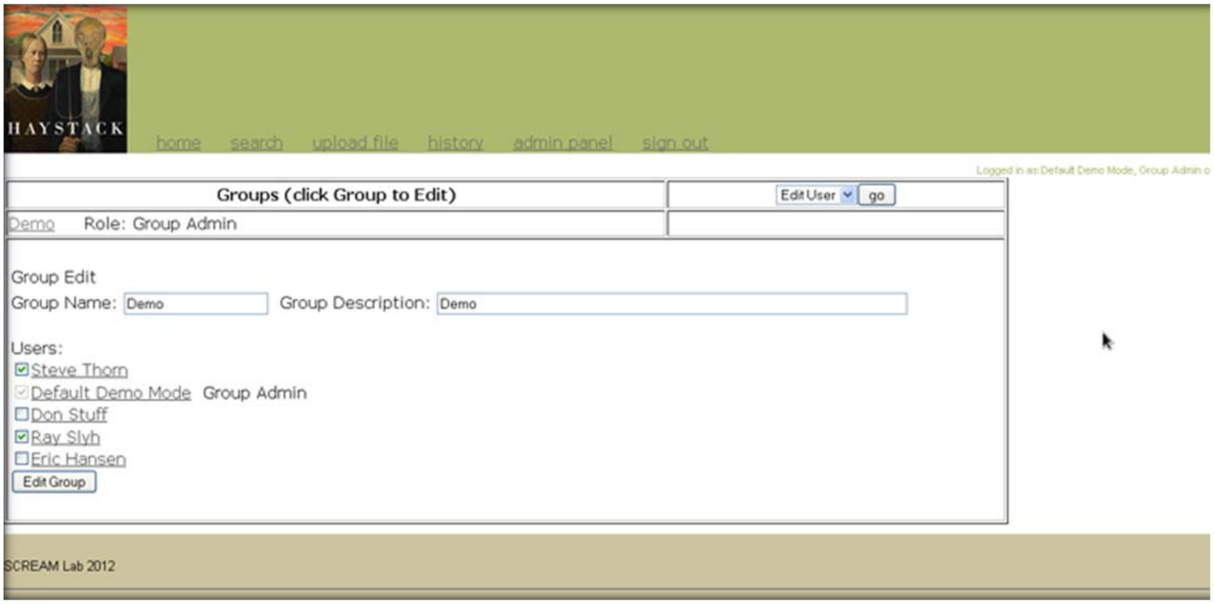
Other experiments that could be run entail partitioning the text that goes into training language models into these different topics and then either adapting or building topic specific language models. In the case of ASR, if transcripts from the first pass recognition are detected to be from a specific topic, a second-pass, topic-specific ASR could be run on those segments to potentially increase the performance. Likewise for statistical MT, domain-specific statistical MT systems could be re-run over specific segments once identified.

### 2.4.4. User Interface (UI) Improvements

This section describes a number of improvements to the user interface of the Haystack MMIER system.

**Administration:** During its initial development phase, the Haystack web interface did not have user authentication or group access controls. Functionality for administrative tasks and permission controls was developed to demonstrate multiuser and multi-group compatibility.

Login scripts, group creation, and hierarchical user/group permissions were created that incorporated a relative database architecture using MySQL and browser session variable controls defined within the Hypertext Preprocessor (PHP) framework. Each group has its own administrator with rights for that specific group. Users can be assigned to multiple groups and have different rights within each group. Groups do not necessarily see what other groups have uploaded to Haystack. Any actions taken by the default “Demo” group are available for all to see at any time. Future designs will implement a tighter control on upload permissions and viewing permissions.



Logged in as:Default Demo Mode, Group Admin o

Groups (click Group to Edit)

Demo Role: Group Admin

Group Edit

Group Name: Demo Group Description: Demo

Users:

☑ Steve Thorn
☑ Default Demo Mode Group Admin
□ Don Stuff
☑ Ray Slyh
□ Eric Hansen

SCREAM Lab 2012

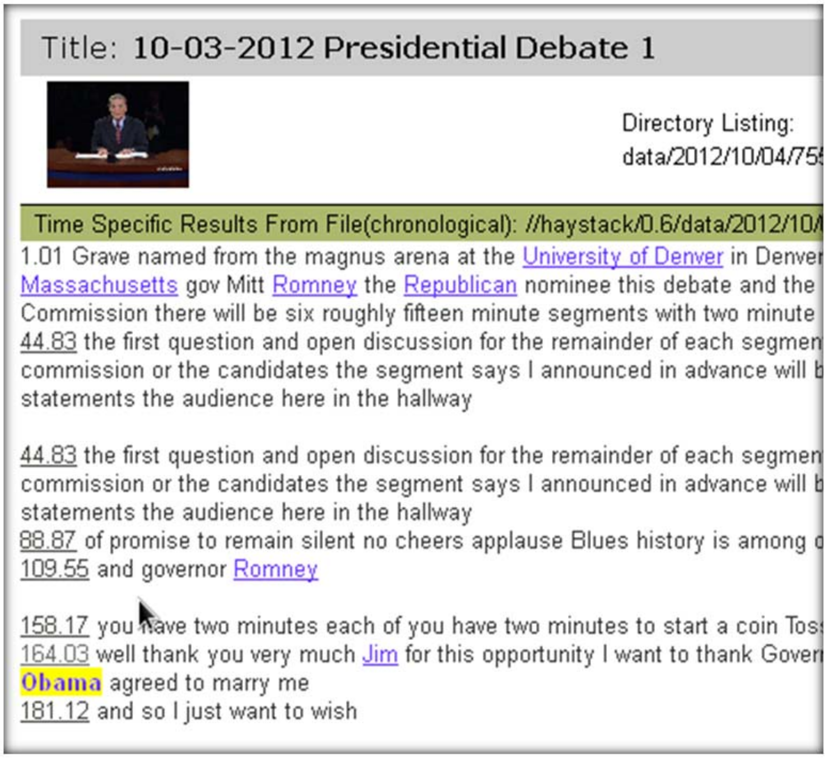
**Figure 3: A Group Administrator Panel Demonstrating how its Modular Design Allows for Editing the Group and its Users in a Single Pass Along with Controls to Delve Deeper into the Editing Process**

A Create, Read, Update, and Delete (CRUD) methodology was adopted for the administrative panels within Haystack. This modification created a modular area for administrators and users alike to edit and self-edit as permission settings allowed. The group administrator panel is shown in Figure 3.

**Search Results:** This section describes changes to the underlying search and retrieval architecture based on Apache Solr as well as user interface changes to support dynamic display of results.

***Solr Update:*** Apache Lucene Solr was updated from Version 1.4 to 3.4. This update allowed for quicker results from a more robust indexing schema. Included in the latest build was an update to Apache Tika, allowing for better extraction from textual documents and a more aggressive extraction process for older Adobe Acrobat Portable Document Format (PDF) files that previously had to be skipped when indexing. Field collapsing was a new feature in this version of Solr, allowing results to be grouped by a selected property and expanded upon when needed.

Lucene Solr incorporates a highlighting function that inserts specific Hyper Text Markup Language (HTML) tags within the returned results. A script was developed to capture the area containing the highlighted section of text, then compare and match it within the complete time-stamped text file in order to find the utterance lines before and after the highlighted section. Each result displays a multiline chunk of text so the user can get a local contextual view of the discussion that was returned. The user can then choose to click on any of the text to be taken immediately to that time within the media file. An example of this display is show in Figure 4.



Title: 10-03-2012 Presidential Debate 1

Directory Listing:
data/2012/10/04/75

Time Specific Results From File(chronological): //haystack/0.6/data/2012/10/

1.01 Grave named from the magnus arena at the University of Denver in Denver
Massachusetts gov Mitt Romney the Republican nominee this debate and the
Commission there will be six roughly fifteen minute segments with two minute
44.83 the first question and open discussion for the remainder of each segmen
commission or the candidates the segment says I announced in advance will b
statements the audience here in the hallway

44.83 the first question and open discussion for the remainder of each segmen
commission or the candidates the segment says I announced in advance will b
statements the audience here in the hallway
88.87 of promise to remain silent no cheers applause Blues history is among c
109.55 and governor Romney

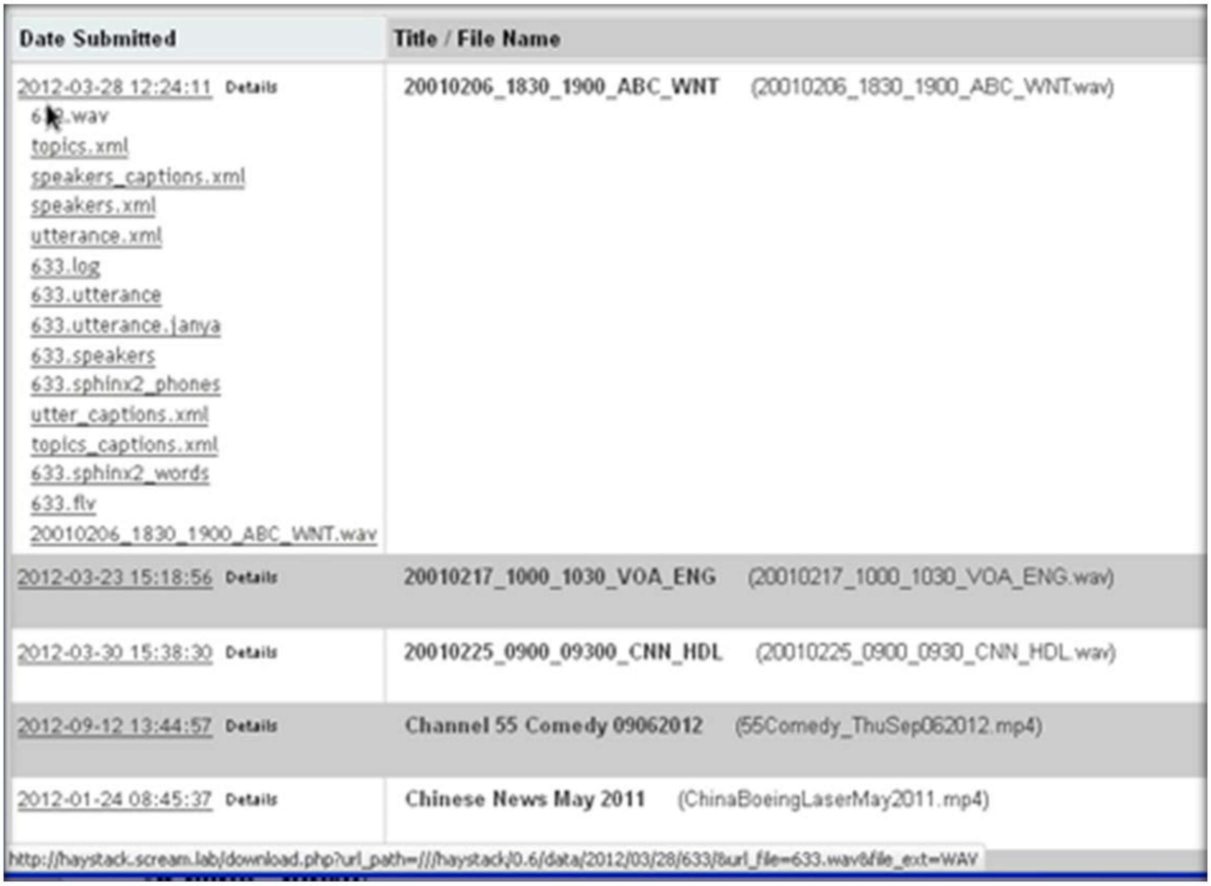
158.17 you have two minutes each of you have two minutes to start a coin Tos
164.03 well thank you very much Jim for this opportunity I want to thank Gover
Obama agreed to marry me
181.12 and so I just want to wish

**Figure 4: The Search Results Page will Display the Highlighted Results but also Display the Utterances Surrounding the Results; A User can Select a Time within the Media File to Begin Playback**

Because each media file might return large amounts of highlighted text from popular terms, this approach to viewing the results could result in substantial amounts of returned text. To combat this issue, the open source JQuery[8] library was adopted for developing some dynamic expansion and collapsing of the text to allow a broad overview of the results and then the ability to drill down into specific results for each media file returned.

***Dynamic Expansion with JQuery:*** JQuery is a free, open source JavaScript library for dynamic update and control of web pages incorporating various features of client-side scripting. Using the standard BubbleInfo class from the JQuery library, the History page was updated to dynamically expand the hyperlinked list of files created during translation with a basic hover-over. The Search and Results pages were combined so that search results would appear dynamically below the query, and the BubbleInfo class was extended for more careful manipulation of the highlighted and segmented results that could then be expanded or collapsed per media file. A list of files in the expanded view is shown in Figure 5. Figure 6 shows a window with search results with one result expanded.

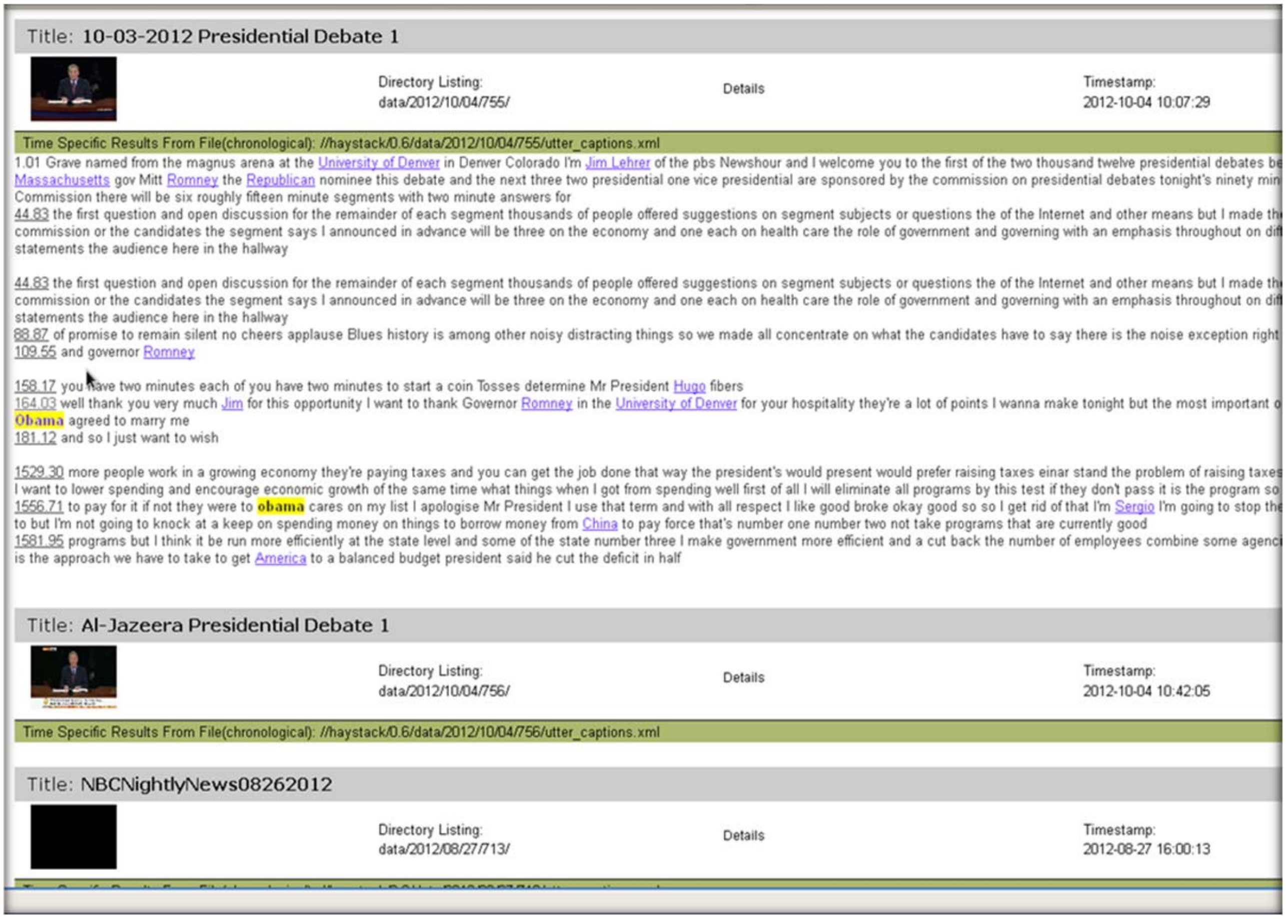
[8] http://jquery.com/



| Date Submitted | Title / File Name |
|---|---|
| 2012-03-28 12:24:11 Details<br>633.wav<br>topics.xml<br>speakers_captions.xml<br>speakers.xml<br>utterance.xml<br>633.log<br>633.utterance<br>633.utterance.janya<br>633.speakers<br>633.sphinx2_phones<br>utter_captions.xml<br>topics_captions.xml<br>633.sphinx2_words<br>633.flv<br>20010206_1830_1900_ABC_WNT.wav | **20010206_1830_1900_ABC_WNT** (20010206_1830_1900_ABC_WNT.wav) |
| 2012-03-23 15:18:56 Details | **20010217_1000_1030_VOA_ENG** (20010217_1000_1030_VOA_ENG.wav) |
| 2012-03-30 15:38:30 Details | **20010225_0900_09300_CNN_HDL** (20010225_0900_0930_CNN_HDL.wav) |
| 2012-09-12 13:44:57 Details | **Channel 55 Comedy 09062012** (55Comedy_ThuSep062012.mp4) |
| 2012-01-24 08:45:37 Details | **Chinese News May 2011** (ChinaBoeingLaserMay2011.mp4) |

http://haystack.scream.lab/download.php?url_path=///haystack/0.6/data/2012/03/28/633/&url_file=633.wav&file_ext=WAV

**Figure 5: Hovering the Mouse Pointer over the Details Tag within the History Page will Expand the Page Dynamically Revealing the Hyperlinked List of Files Created During the Translation Process**



Title: 10-03-2012 Presidential Debate 1

Directory Listing: data/2012/10/04/755/

Details

Timestamp: 2012-10-04 10:07:29

Time Specific Results From File(chronological): //haystack/0.6/data/2012/10/04/755/utter_captions.xml

1.01 Grave named from the magnus arena at the University of Denver in Denver Colorado I'm Jim Lehrer of the pbs Newshour and I welcome you to the first of the two thousand twelve presidential debates be
Massachusetts gov Mitt Romney the Republican nominee this debate and the next three two presidential one vice presidential are sponsored by the commission on presidential debates tonight's ninety min
Commission there will be six roughly fifteen minute segments with two minute answers for
44.83 the first question and open discussion for the remainder of each segment thousands of people offered suggestions on segment subjects or questions the of the Internet and other means but I made th
commission or the candidates the segment says I announced in advance will be three on the economy and one each on health care the role of government and governing with an emphasis throughout on dif
statements the audience here in the hallway

44.83 the first question and open discussion for the remainder of each segment thousands of people offered suggestions on segment subjects or questions the of the Internet and other means but I made th
commission or the candidates the segment says I announced in advance will be three on the economy and one each on health care the role of government and governing with an emphasis throughout on dif
statements the audience here in the hallway
88.87 of promise to remain silent no cheers applause Blues history is among other noisy distracting things so we made all concentrate on what the candidates have to say there is the noise exception right
109.55 and governor Romney

158.17 you have two minutes each of you have two minutes to start a coin Tosses determine Mr President Hugo fibers
164.03 well thank you very much Jim for this opportunity I want to thank Governor Romney in the University of Denver for your hospitality they're a lot of points I wanna make tonight but the most important o
Obama agreed to marry me
181.12 and so I just want to wish

1529.30 more people work in a growing economy they're paying taxes and you can get the job done that way the president's would present would prefer raising taxes einar stand the problem of raising taxes
I want to lower spending and encourage economic growth of the same time what things when I got from spending well first of all I will eliminate all programs by this test if they don't pass it is the program so
1556.71 to pay for it if not they were to obama cares on my list I apologise Mr President I use that term and with all respect I like good broke okay good so so I get rid of that I'm Sergio I'm going to stop the
to but I'm not going to knock at a keep on spending money on things to borrow money from China to pay force that's number one number two not take programs that are currently good
1581.95 programs but I think it be run more efficiently at the state level and some of the state number three I make government more efficient and a cut back the number of employees combine some agenci
is the approach we have to take to get America to a balanced budget president said he cut the deficit in half

Title: Al-Jazeera Presidential Debate 1

Directory Listing: data/2012/10/04/756/

Details

Timestamp: 2012-10-04 10:42:05

Time Specific Results From File(chronological): //haystack/0.6/data/2012/10/04/756/utter_captions.xml

Title: NBCNightlyNews08262012

Directory Listing: data/2012/08/27/713/

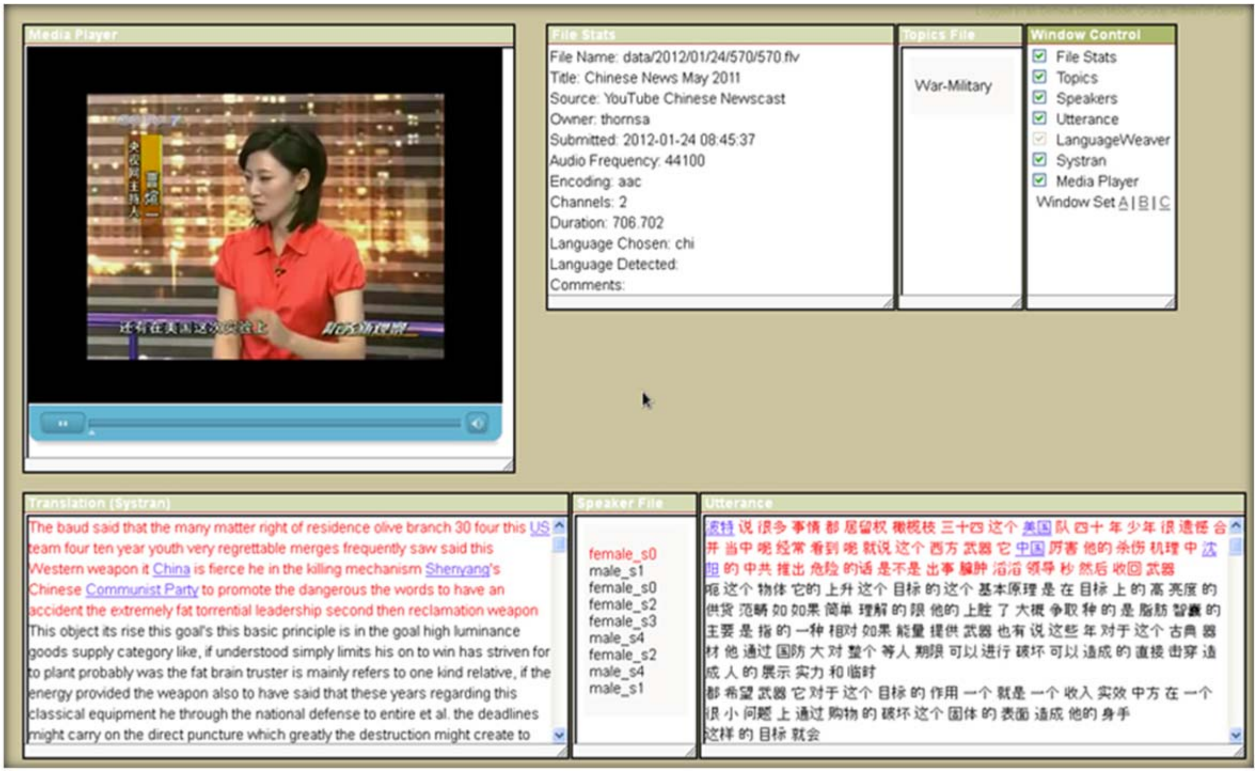
Details

Timestamp: 2012-08-27 16:00:13

**Figure 6: By Incorporating and Extending the JQuery Library, the Search Results are more Easily Manageable through Expanding and Collapsing Sections of the Highlighted Utterances**

**Media Player**: This section describes changes to the media player pages that added dynamic windows for displaying ASR transcripts, translations, topics, *etc*. These windows were dynamic in terms of their placement on the page, and their content was made to be dynamic in that the content could scroll in sync with the playing of the audio.

***Dynamic Window Placement:*** The original Haystack interface for interacting with a media file became very cumbersome rather quickly. The interface for the media file playback consisted of Topic Detection, Speaker Identification, Utterance and Translated captioning, and the Meta properties of the file itself. There were also various mutations that had to be kept in mind because there were modules for video with or without translation, audio with or without translation, and translated text documents. To improve the interface, code was developed to make dynamic windows with some static properties**.**



File Name: data/2012/01/24/570/570.flv
Title: Chinese News May 2011
Source: YouTube Chinese Newscast
Owner: thornsa
Submitted: 2012-01-24 08:45:37
Audio Frequency: 44100
Encoding: aac
Channels: 2
Duration: 706.702
Language Chosen: chi
Language Detected:
Comments:

War-Military

☑ File Stats
☑ Topics
☑ Speakers
☑ Utterance
☑ LanguageWeaver
☑ Systran
☑ Media Player
Window Set A | B | C

The baud said that the many matter right of residence olive branch 30 four this US team four ten year youth very regrettable merges frequently saw said this Western weapon it China is fierce he in the killing mechanism Shenyang's Chinese Communist Party to promote the dangerous the words to have an accident the extremely fat torrential leadership second then reclamation weapon
This object its rise this goal's this basic principle is in the goal high luminance goods supply category like, if understood simply limits his on to win has striven for to plant probably was the fat brain truster is mainly refers to one kind relative, if the energy provided the weapon also to have said that these years regarding this classical equipment he through the national defense to entire et al. the deadlines might carry on the direct puncture which greatly the destruction might create to

female_s0
male_s1
female_s0
female_s2
female_s3
male_s4
female_s2
male_s4
male_s1

波特 说 很多 事情 都 居留权 橄榄枝 三十四 这个 美国 队 四十 年 少年 很 遗憾 合并 当中 呢 经常 看到 呢 就说 这个 西方 武器 它 中国 厉害 他的 杀伤 机理 中 沈阳 的 中共 推出 危险 的话 是不是 出事 臃肿 滔滔 领导 秒 然后 收回 武器
呃 这个 物体 它的 上升 这个 目标 的 这个 基本原理 是 在 目标 上 的 高 亮度 的 供货 范畴 如 如果 简单 理解 的 限 他的 上胜 了 大概 争取 种 的 是 脂肪 智囊 的 主要 是 指 的 一种 相对 如果 能量 提供 武器 也有 说 这些 年 对于 这个 古典 器材 他 通过 国防 大 对 整个 等人 期限 可以 进行 破坏 可以 造成 的 直接 击穿 造成 人 的 展示 实力 和 临时
都 希望 武器 它 对于 这个 目标 的 作用 一个 就是 一个 收入 实效 中方 在 一个 很 小 问题 上 通过 购物 的 破坏 这个 固体 的 表面 造成 他的 身手
这样 的 目标 就会

```
<script type="text/javascript">

var flashwin=window.open("flashbox",
"iframe_MMPlayerUtter.php?fileName=data/2012/10/18/765/765.flv&pathName=data/2012/10/18/765/&short_vid=0&language=eng", "Media Player",
"width=495px,height=415px,left=20px,top=20px,resize=1,scrolling=0","recal")

var speakerswin=window.open("speakerbox",
"iframe_MMspeakers.php?fileName=data/2012/10/18/765/765.flv&pathName=data/2012/10/18/765/&speakers_path=../data/2012/10/18/765/speakers_captions.xml","Speaker File","width=125px,height=250px,left=1035px,top=20px,resize=1,scrolling=1","recal");

var topicswin=window.open("topicsbox",
"iframe_MMtopics.php?fileName=data/2012/10/18/765/765.flv&pathName=data/2012/10/18/765/&topics_path=../data/2012/10/18/765/topics_captions.xml", "Topics File",
"width=125px,height=250px,left=905px,top=20px,resize=1,scrolling=1", "recal")

var utterwin=window.open("utterbox",
"iframe_MMutter_solo.php?fileName=data/2012/10/18/765/765.flv&pathName=data/2012/10/18/765/&language=eng", "Utterance",
"width=765px,height=350px,left=550px,top=315px,resize=1,scrolling=1", "recal")

</script>
```

**Figure 7: Haystack Media Player Page Incorporating the JavaScript-Driven Windows for Placement; the Corresponding HTML is shown below the Page**

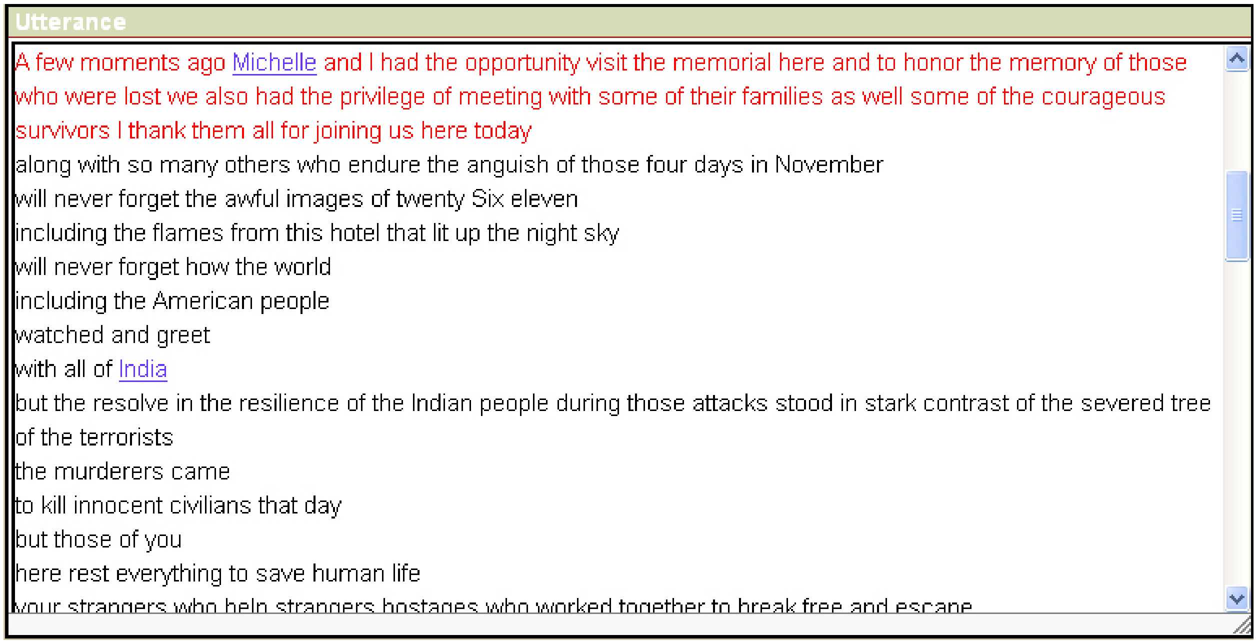
Dynamic windows developed in JavaScript, as shown in Figure 7, were created with certain static properties, so that the windows could now be moved around, layered, or hidden according to the user's wishes. Preset values for the windows were created pertaining to the layout of the page specific to whether the media file was audio or video and whether it was translated or contained only ASR transcripts with no translations.

***Dynamic Captioning:*** Adding dynamic captioning to the processed and translated media files was a way to gain control of the sometimes lengthy translation and utterance text files. The overall concept for the captioning consisted of scrollable text boxes, where the currently spoken text was highlighted and would move along in time (i.e., in sync) with the media file as it was played.

Because of the ASR and MT processes, the transcripts and translations were already broken into time-stamped utterances, so the text was transformed into an Extensible Markup Language (XML) document for better manipulation in Flash and JavaScript. The text would be displayed in HTML format within the text boxes to enable hyperlinks and to take advantage of the Anchor and Bookmark tags.

Adobe Flash has a built-in ExternalInterface mechanism for dynamically communicating back and forth with JavaScript on a web page. When a media file in Haystack is played, it accesses the Flash FLVPlayer component for display that in turn sends out timed event notices firing the ExternalInterface to an awaiting JavaScript. The JavaScript polls the text caption timestamps and discovers what lines of text are now current. It then highlights the text by changing its style in the HTML and calls it as an active Bookmark that maneuvers it to the top of the scroll box.

Also, the lines of captioned text have the timestamp hyperlinked so that at any time a user could scroll within the text and click on it, firing the JavaScript, returning back through the



Utterance

A few moments ago Michelle and I had the opportunity visit the memorial here and to honor the memory of those who were lost we also had the privilege of meeting with some of their families as well some of the courageous survivors I thank them all for joining us here today
along with so many others who endure the anguish of those four days in November
will never forget the awful images of twenty Six eleven
including the flames from this hotel that lit up the night sky
will never forget how the world
including the American people
watched and greet
with all of India
but the resolve in the resilience of the Indian people during those attacks stood in stark contrast of the severed tree
of the terrorists
the murderers came
to kill innocent civilians that day
but those of you
here rest everything to save human life
your strangers who help strangers hostages who worked together to break free and escape

**Figure 8: An Example of the Utterance Captioning Window with Highlighted Text as it is Positioned to the Top of the Scrolling Window**

ExternalInterface method, and triggering a “seek,” scrubbing to that exact time within the media file and playing from there.

**Moses Server:** For the initial steps of incorporating Moses[9] translation into the Haystack system, a standalone version of Moses Server was run on a basic system with access to two different language models - namely, a Europarl model and a model trained on United Nations text. Each translation pair and direction (*e.g.*, French-to-English or English-to-French) was assigned specific port numbers so that specifically designed WSDLs could be called upon as needed.

```
./mosesserver -config
/haystack/0.6/htdocs/services/lang_ini/fren_europarl/moses.ini --server-port 8081
--server-log /home/log.txt
```

**Figure 10: Command Line Startup for Moses Server Assigning the French-to-English Europarl Translation Service to Port 8081**

The web services for Moses translation fire a PERL-based Remote Procedure Call (RPC) client for communicating with the Moses Server. The translation is executed as a line-by-line translation and stitched back together for incorporation into the caption windows.

[9] www.statmt.org/moses/

## 3.0 CONCLUSIONS

In conclusion, work has been accomplished in the areas of ASR, MT, and topic detection, especially in the context of an in-house multilingual multimedia information extraction and retrieval system called Haystack.

For ASR, systems were developed for several languages and integrated into the Haystack system. Improved performance was shown by interpolating in-domain LMs with LMs built on additional out-of-domain text. Several experiments were conducted on methods to reduce the effects of OOV words encountered by an ASR system; however, these experiments generally showed only small improvements. Systems based on the graphone method of Bisani and Ney were developed for both English and Croatian, with greater improvement found for Croatian than for English due to the greater morphological complexity of Croatian. However, for both languages, the WERs were generally improved more by expanding the number of words in the pronunciation dictionaries and LMs. Some small improvements were obtained for Arabic ASR by segmenting words into morphemes using a system called MADA and building sub-word LMs based on the morphemes. The performance was generally best when only the least occurring words were segmented. A similar result was found for Russian ASR, where segmenting the least occurring words using a system based on Morfessor or on a Russian stemmer and building sub-word LMs improved WERs a small amount. Overall, there is much more work to be done in addressing the OOV problem as the gains achieved so far were small. Finally, multiple ASR systems were developed for the 2012 IWSLT ASR evaluation, but at the time of this report, the final evaluation results were not known.

For MT, systems were developed for the 2012 IWSLT AE evaluation, with some improve performance found by segmenting Arabic into morphemes, but keeping the definite article (Al-) attached to the word stems a opposed to separating it. Again, at the time of this report, the final evaluation results were not known.

For topic detection, LMs and SVMs were investigated, and it was found that SVMs using the normalized tf-idf feature vectors generally performed better than either LMs or SVMs using the normalized tf feature vectors. However, when these detectors were considered for the Haystack system, it was found that they did not perform as well on the types of data that the Haystack system was being used for. Therefore, a method based on counting topic-specific keywords was considered and found to be surprising useful. This system is easy to implement, fast in execution, and can be ported to new languages relatively easy by translating the topic-specific keywords.

Finally, several modifications were made to the Haystack MMIER system. In particular, improvements were made to the processing pipeline, ASR systems were added, the underlying search and retrieval architecture was upgraded, and the user interface was improved.

## 4.0 REFERENCES

1. M. Bisani and H. Ney, “Open Vocabulary Speech Recognition with Flat Hybrid Models,” in *Proc. of Interspeech*, Istanbul, Turkey, 2005.
2. M. Bisani and H. Ney, “Joint-Sequence Models for Grapheme-to-Phoneme Conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.
3. D. Graff *et al.*, “CSR-III Text,” *Linguistic Data Consortium*, Philadelphia, 1995 (Available at http://www.ldc.upenn.edu).
4. A. Stolcke, “SRILM – An Extensible Language Modeling Toolkit,” in *Proc. of the International Conference on Spoken Language Processing,* Denve CO, 2002.
5. J. Garofolo *et al.*, “CSR-I (WSJ0) Complete,” *Linguistic Data Consortium*, Philadelphia, 2007 (Available at http://www.ldc.upenn.edu).
6. “CSR-II (WSJ1) Complete,” *Linguistic Data Consortium*, Philadelphia PA, 1994 (Available at http://www.ldc.upenn.edu).
7. Cambridge University Engineering Department, The HTK Book, 2009 (Available at http://htk.eng.cam.ac.uk).
8. T. Schultz, “GlobalPhone: a Multilingual Speech and Text Database Developed at Karlsruhe University,” in *Proc. of the International Conference on Spoken Language Processing*, Denver CO, 2002.
9. H. Nizar *et al.*, “MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization,” in *Proc. of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt, 2009.
10. S. Strassel *et al.*, “TDT4 Multilingual Text and Annotations,” *Linguistic Data Consortium*, Philadelphia, 2005 (Available at http://www.ldc.upenn.edu).
11. D. Graff *et al.*, “TDT5 Multilingual Text,” *Linguistic Data Consortium*, Philadelphia PA, 2006 (Available at http://www.ldc.upenn.edu).
12. A. El-Desoky *et al.*, “Investigating the use of Morphological Decomposition and Diacritization for Improving Arabic LVCSR,” in *Proc. of Interspeech*, Brighton, U.K., 2009.
13. Mathias Creutz and Krista Lagus, “Inducing the Morphological Lexicon of a Natural Language from Unannotated Text,” in *Proc. of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Espoo, Finland, 2005.
14. J. Tiedemann, “Parallel Data, Tools and Interfaces in OPUS,” in *Proc. of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012.
15. A. Aminzadeh *et al.*, “The MIT-LL/AFRL IWSLT-2009 MT System,” in *Proc. of the International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009.
16. Topic Detection and Tracking: http://projects.ldc.upenn.edu/TDT/.
17. M.F. Porter, “An Algorithm for Suffix Stripping,” *Program*, 14(3), pp 130-137, 1980.
18. P.R. Clarkson and R. Rosenfeld, “Statistical Language Modeling using the CMU-Cambridge Toolkit,” in *Proc. of Eurospeech,* Rhodes, Greece, 1997.

19. R. Collobert and S. Bengio, "SVMTorch: Support Vector Machines for Large-Scale Regression Problems," *J. Machine Learning Research,* vol. 1, pp. 143-160, 2001.
20. C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology,* 2:27:1-27:27, 2011.
21. C. Manning, P. Raghaven, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.
22. M. Saraçlar, "Turkish Broadcast News Speech and Transcripts," *Linguistic Data Consortium*, Philadelphia, 2012 (Available at http://www.ldc.upenn.edu).
23. M. Pilevar *et al*., "TEP: Tehran English-Persian Parallel Corpus," in *Proc. of the 12th International Conference on Intelligent Text Processing and Computational Linguists*, Tokyo, Japan, February 2011.
24. D. Graff *et al*., "French Gigaword Third Edition," *Linguistic Data Consortium*, Philadelphia PA, 2011 (Available at http://www.ldc.upenn.edu).
25. D. Graff, "European Language Newspaper Text," *Linguistic Data Consortium*, Philadelphia PA, 1995 (Available at http://www.ldc.upenn.edu).
26. D. Santos and P. Rocha, "CETEMPúblico," *Linguistic Data Consortium*, Philadelphia PA, 2001 (Available at http://www.ldc.upenn.edu).
27. J. Wright and D. Graff, "Portuguese Newswire Text," *Linguistic Data Consortium*, Philadelphia PA, 1999 (Available at http://www.ldc.upenn.edu).
28. A. Mendonça *et al*., "Spanish Gigaword Third Edition," *Linguistic Data Consortium*, Philadelphia PA, 2011 (Available at http://www.ldc.upenn.edu).
29. E. Munoz, "1997 Spanish Broadcast News Transcripts (HUB4-NE)," *Linguistic Data Consortium*, Philadelphia PA, 1998 (Available at http://www.ldc.upenn.edu).
30. A. R. Aminzadeh *et al*., "The MIT-LL/AFRL IWSLT-2011 MT System," in *Proc. of the International Workshop on Spoken Language Translation*, San Francisco CA, 2011.
31. R. Parker *et al*., "English Gigaword Fifth Edition," *Linguistic Data Consortium*, Philadelphia PA, 2011 (Available at http://www.ldc.upenn.edu).
32. R. Moore and W. Lewis, "Intelligent Selection of Language Model Training Data," in *Proc. of the Association of Computational Linguistics Conference*, Uppsala, Sweden, 2010.
33. T. Mikolov *et al*., "Strategies for Training Large Scale Neural Network Language Models," in *Proc. of the Automatic Speech Recognition and Understanding Workshop*, Hawaii, 2011.
34. W. Shen, B. Delaney, T. Anderson, and R. Slyh, "The MIT-LL/AFRL IWSLT-2007 MT System," in *Proc. of the International Workshop on Spoken Language Translation*, Trento, Italy, 2007.
35. W. Shen, B. Delaney, T. Anderson, and R. Slyh, "The MIT-LL/AFRL IWSLT-2008 MT System," in *Proc. of the International Workshop on Spoken Language Translation*, Waikiki HI, 2008.
36. W. Shen, T. Anderson, R. Slyh, and A. R. Aminzadeh, "The MIT-LL/AFRL IWSLT-2010 MT System," in *Proc. of the International Workshop on Spoken Language Translation*, Paris, France, 2010.

37. R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin, “Arabic Morphological Tagging, Diacritization, and Lemmatization using Lexeme Models and Feature Ranking,” in *Proc. of ACL-08: HLT Short Papers,* Columbus OH, 2008.
38. J. Wuebker, M. Huck, S. Mansour, M. Freitag, M. Feng, S. Peitz, C. Schmidt, and H. Ney, “The RWTH Aachen Machine Translation System for IWSLT 2011,” in *Proc. of the International Workshop on Spoken Language Translation*, San Francisco CA, 2011.
39. G. Foster, R. Kuhn, and H. Johnson, “Phrasetable Smoothing for Statistical Machine Translation,” in *Proc. of EMNLP,* Sydney, Australia, 2006.

## LIST OF ACRONYMS & GLOSSARY

| | |
|---|---|
| AE | Arabic-to-English |
| ARPA | Advanced Research Projects Agency |
| AFRL/RHXS | Air Force Research Laboratory/Human Effectiveness Directorate, Anticipate & Influence Behavior Division, Sense-making & Organizational Effectiveness Branch |
| AM | Acoustic Model |
| Apache | A group that supports open source software development projects |
| ARB | Arabic |
| ASR | Automatic Speech Recognition |
| ATB | Arabic Treebank |
| BLEU | Bilingual Evaluation Understudy |
| CMLLR | Constrained Maximum Likelihood Linear Regression |
| CMU | Carnegie Mellon University |
| CMU-LMTK | Carnegie Mellon University Language Modeling ToolKit |
| CoMMA | Count Mediated Morphological Analysis |
| CRUD | Create, Read, Update & Delete |
| CPAN | Comprehensive Perl Archive Network |
| CSR | Continuous Speech Recognition |
| DER | Diarization Error Rate |
| DET | Detection Error Trade-Off |
| EER | Equal Error Rate |
| ENG | English |
| FLARe | Foreign Language Analysis and Recognition |
| FLVPlayer | Flash Video Player |
| G2P | Grapheme to Phoneme |
| GALE | Global Autonomous Language Exploitation |
| GlobalPhone | A multilingual text and speech database |
| Haystack | An internal SCREAM Lab project to integrate the various SCREAM Lab capabilities into a system to index, analyze, translate, store and retrieve multilingual information from rich multimedia documents in various languages |
| HDecode | Cambridge University large vocabulary continuous speech recognizer |
| HLDA | Heteroscedastic Linear Discriminate Analysis |
| HMM | Hidden Markov Model |

| | |
|---|---|
| HPW | Human Performance Wing |
| HTML | HyperText Markup Language |
| HTK | Hidden Markov Model Toolkit |
| IR | Information Retrieval |
| IWSLT | International Workshop on Spoken Language Translation |
| Java | Refers to a number of computer software products and specifications from Oracle that together provide a system for developing application software and deploying it in a cross-platform environment |
| JavaScript | JavaScript is a script language typically used to enable programmatic access to computational objects within a host environment, commonly a web browser |
| JQuery | Open source JavaScript library for dynamic update and control of web pages |
| KN | Kneser-Ney |
| Language Weaver | A commercial machine translation system |
| LER | Letter Error Rate |
| LIBSVM | A library for support vector machines |
| LM | Language Model |
| Lucene | A high-performance, full-featured text search engine library written entirely in Java |
| LVSCR | Large Vocabulary Continuous Speech Recognizer |
| MAN | Mandarin Chinese |
| MERT | Minimum Error Rate Training |
| MFCC | Mel-Frequency Cepstral Coefficient |
| MMIER | Multilingual Multimedia Information Extraction & Retrieval |
| Morfessor | Software developed at the Helsinki University of Technology for unsupervised learning of morphology |
| Moses | A statistical machine translation system |
| MPE | Minimum Phone Error |
| MT | Machine Translation (a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another) |
| MySQL | A relational database management system that runs as a server providing multi-user access to a number of databases |
| NE | Named Entity |
| NLP | Natural Language Processing |

| | |
|---|---|
| OOV | Out-of-Vocabulary |
| PDF | Portable Document Format (Adobe) |
| PER | Phoneme Error Rate |
| Perl | A high-level, general-purpose, interpreted, dynamic programming language |
| PHP | Hypertext Preprocessor (a widely used, general-purpose scripting language that was originally designed for web development to produce dynamic web pages) |
| PLP | Perceptual Linear Prediction |
| RNNLM | Recurrent Neural Network Language Model |
| RNNME | Recurrent Neural Network Maximum Entropy |
| SAT | Speaker Adaptive Training |
| SCREAM | Speech and Communication Research, Engineering, Analysis, and Modeling |
| Sequitur G2P | Software developed at RWTH Aachen University for training grapheme to phoneme systems |
| Solr | An open source enterprise search platform from the Apache Lucene project |
| SRILM | Stanford Research Institute Language Modeling; A language modeling toolkit developed by Stanford Research Institute |
| SVM | Support Vector Machine |
| Systran | A commercial machine translation system |
| TDT | Topic Detection and Tracking corpus |
| TED | Technology, Entertainment, and Design |
| TEP | Tehran English-Persian parallel corpus |
| Tika | A toolkit from the Apache Software Foundation for detecting and extracting metadata and structured text content from various documents using existing parser libraries |
| Torch | A Matlab-like environment for machine learning algorithms |
| TRANSTAC | Translation System for Tactical Use corpus |
| UI | User Interface |
| WER | Word Error Rate |
| Wikipedia | A free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation |
| WSDL | Web Service Definition Language |
| WSJ | Wall Street Journal |
| XML | eXtensible Markup Language |